HIV Database Workshop

www.hiv.lanl.gov

seq-info@lanl.gov

Presenters: Brian Gaschen, Bette Korber, Thomas Leitner, Brian Foley, Karina Yusim

Database staff: Werner Abfalterer, Chuck Calef, Robert Funkhouser, Kristina Kommander, Dorothy Lang, John Mokili, Una Smith, Jim Szinger, Ming Zhang

Regular Contributors: John Mellors. Robert Shafer, and Christian Brander
Project Officer: James Bradac, NIAID, NIH

Carla Kuiken: www.hcv.lanl.gov

Theoretical Biology and Biophysics, T-10
Los Alamos National Laboratory

Workshop Topics

Introduction - 10 min

Sequence Database

Brian Gaschen - 30 min

Basic sequence search interface and on-the-fly alignments
Geography search interface
GeneCutter - processing nucleotide sequences
N-glycosite - infer N-linked glycosylation (if time permits)

Thomas Leitner - 30 min

HIV database alignments and subtype reference sequences
Comparing "new" sequences with our reference sequences
Build a simple neighbor joining tree through the web
Using the new RIP tool for recombination analysis

Brian Foley - 10 min

3D views of HIV macromolecular structures

Break - 15 min

Immunology Database

Bette Korber - 30 min

HIV/SIV sequence locator tool
CTL search page
Ab search page
Epitope maps

Karina Yusim - 30 min

Peptgen - list peptides for reagent development
Motifscan - find HLA anchor residues in a protein sequence
ELF - epitope location finder

Vaccine Trial Database

Brian Foley - 25 min

SHIV maps
SIV alignments
Vaccine database searches

About the instructors

- **Bette Korber** is the senior biologist on the Los Alamos HIV database project, and chief editor of the HIV immunology database.

- **Thomas Leitner** has recently replaced Dr. Carla Kuiken as chief editor for the HIV sequence database. Dr. Leitner brings to the project a decade of experience in the field of HIV evolution and sequence analysis.

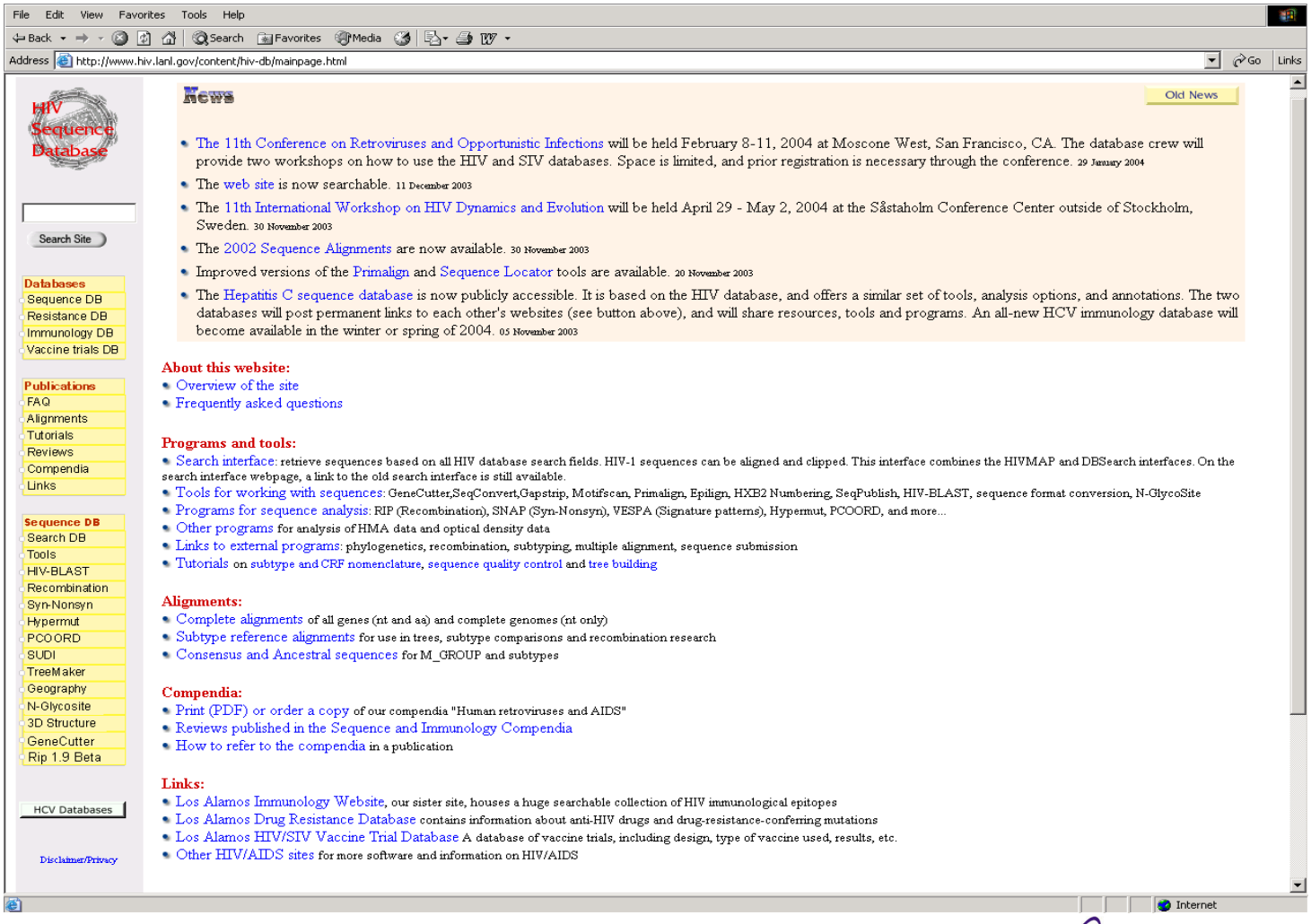
- **Brian Gaschen** is the head programmer of the HIV sequence database project, and developed the HIV sequence relational database, as well as the code for many of the search and analysis tools we will be demonstrating.

- **Karina Yusim** is a postdoctoral fellow who has been involved in analysis of the data included in both the HIV sequence and immunology databases.

- **Brian Foley** joined the HIV Database team in 1995 and has extensive experience with multiple sequence alignments, phylogenetic analyses, epidemiology, the vaccine database and analyses of protein 3D structures.

Workshop Goals

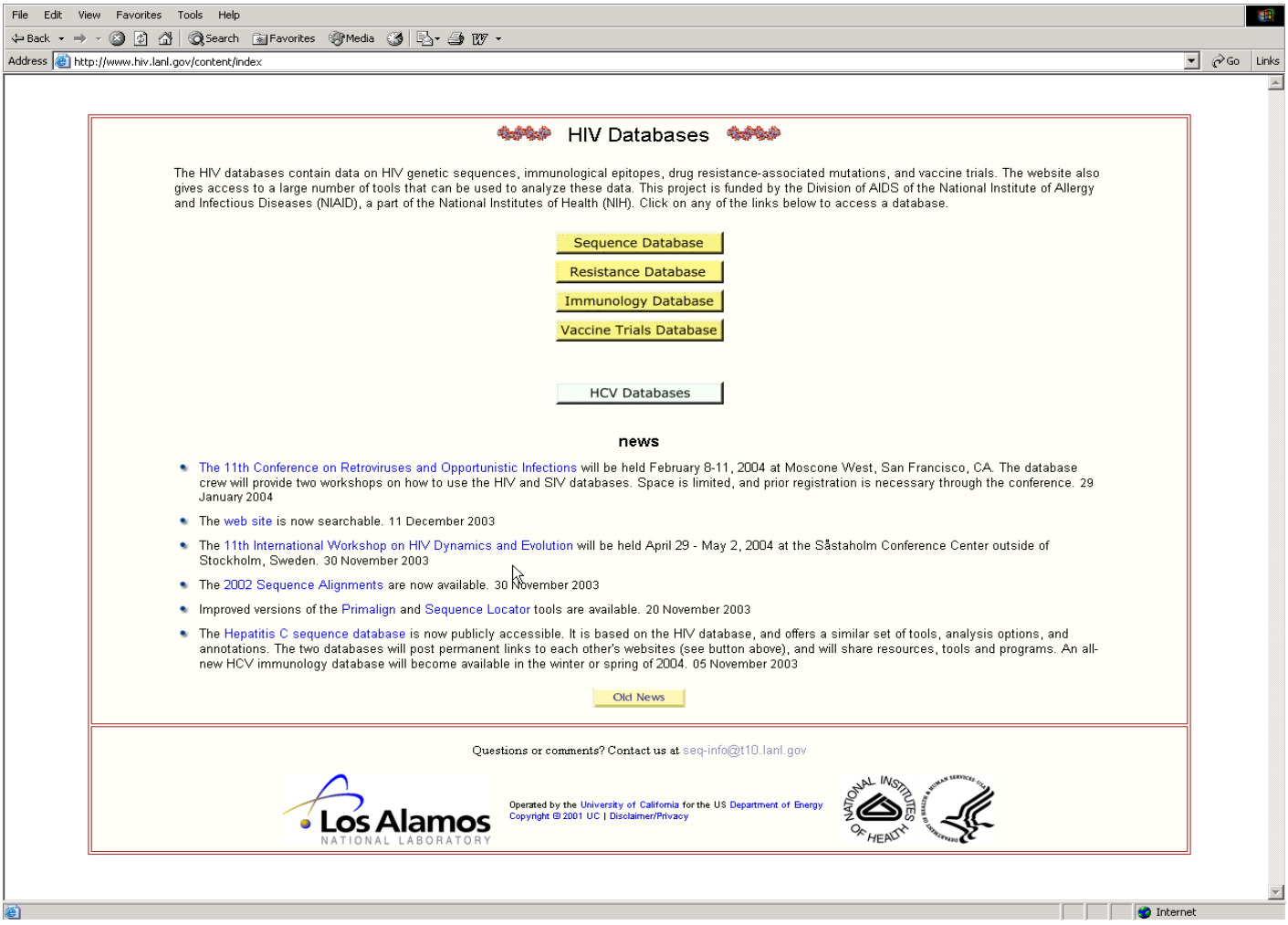
- Understanding the database content, how information was obtained, and what is available
- Database searching
- Quality control tools
- Tools for analyses



Los Alamos NATIONAL LABORATORY

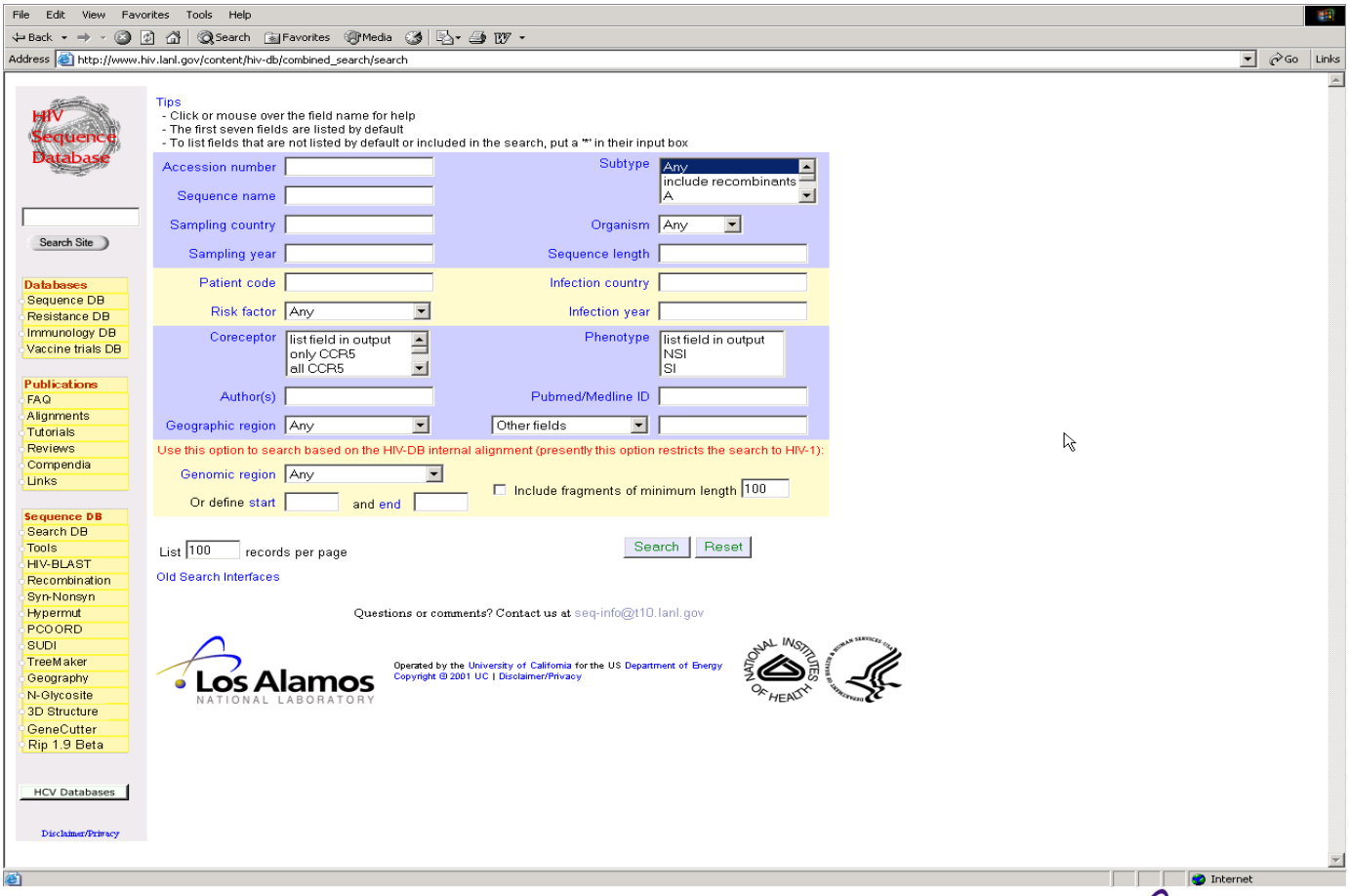
The HIV databases

- HIV Sequence database – founded 1986, G. Myers
  - Relational database, data from GenBank with added fields from the literature
  - Alignments – align indels and reduce multiple sequences per person
  - Annual hard copy and reviews
  - Web search interfaces: subtype, phenotype, geographic, sampling year…
  - Analysis tools
- HIV Immunology database – founded 1995, B. Korber
  - Comprehensive HIV epitope database , 300-400 papers a year
  - Integrate HIV immunological and sequence data
  - Annual hard copy and reviews
  - Web search interfaces: epitope, protein, HLA type, immunogen, keywords
  - Analysis tools for immunologists
- HIV Drug Resistance database, founded 1997, J. Mellors
  - A searchable web listing of drug resistance mutations and literature links, updated annually by Dr. Mellors
- HIV Vaccine database, founded 2003, J. Mokili
  - A searchable relational database of published primate vaccine trials



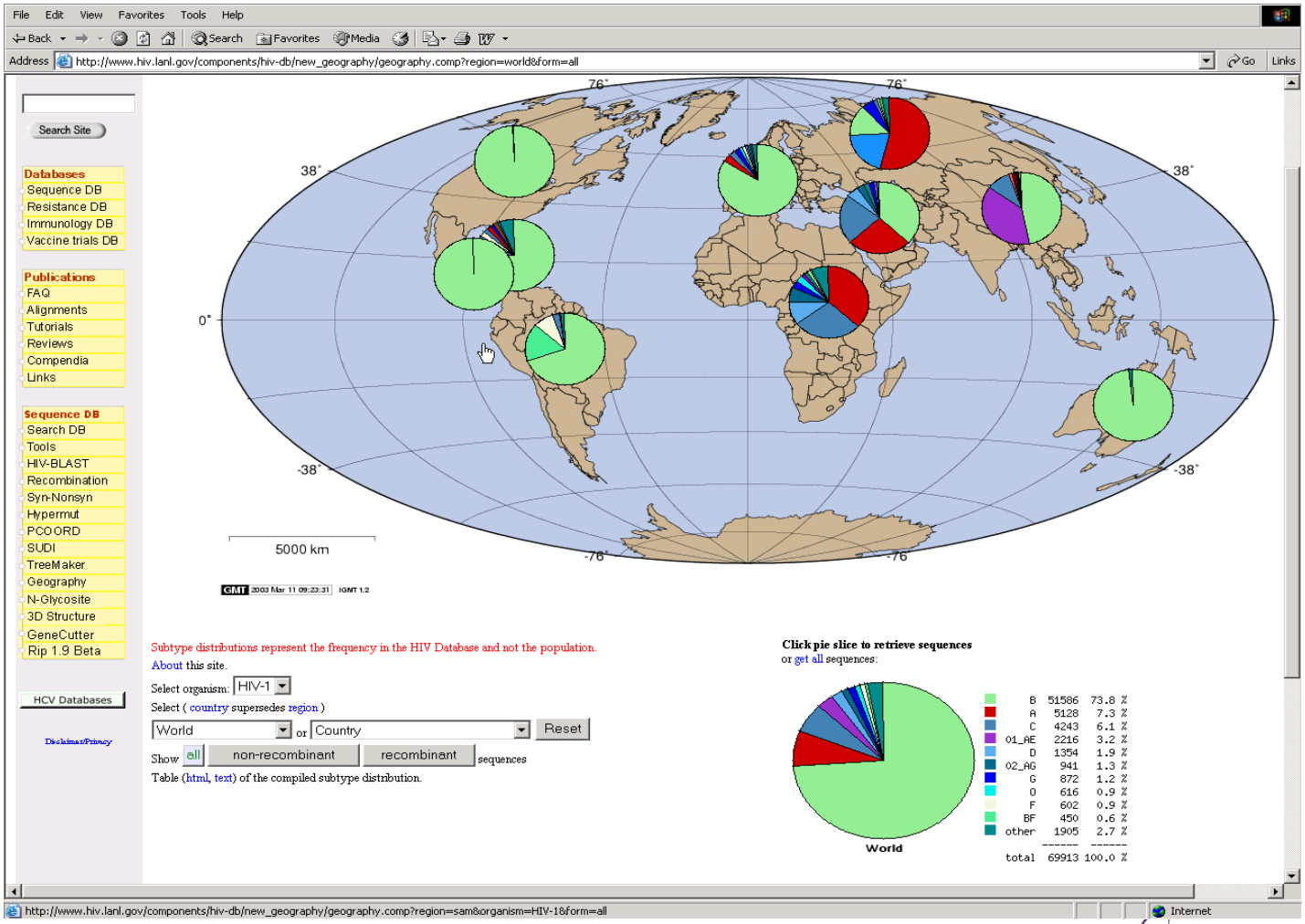
Search Interface

- **Help**
  - Tips at the top of the page are often overlooked
    - Ranges, operators, wildcards, logical groupings
  - Field names are clickable, also mouse-overs
    - Example: “Sampling country” gives two-letter ISO country codes
- **Searches**
  - Searches are case-insensitive
  - Records are searchable through sequence, patient, genomic region, or publication information
  - First seven fields will appear in search results page by default
  - A “*” in a textbox will cause that field to be included in the results page
  - Patient information (Infection year, Infection country) is different than sequence information (Sampling year and Sampling country)
- **Results**
  - Can select not aligned, or aligned based on multiple pair wise alignments – alignments are good, but still need hand editing for an optimal alignment
  - Select all or a subset of sequences for download
  - Sequences can be re-ordered by clicking on fields at the top of the page



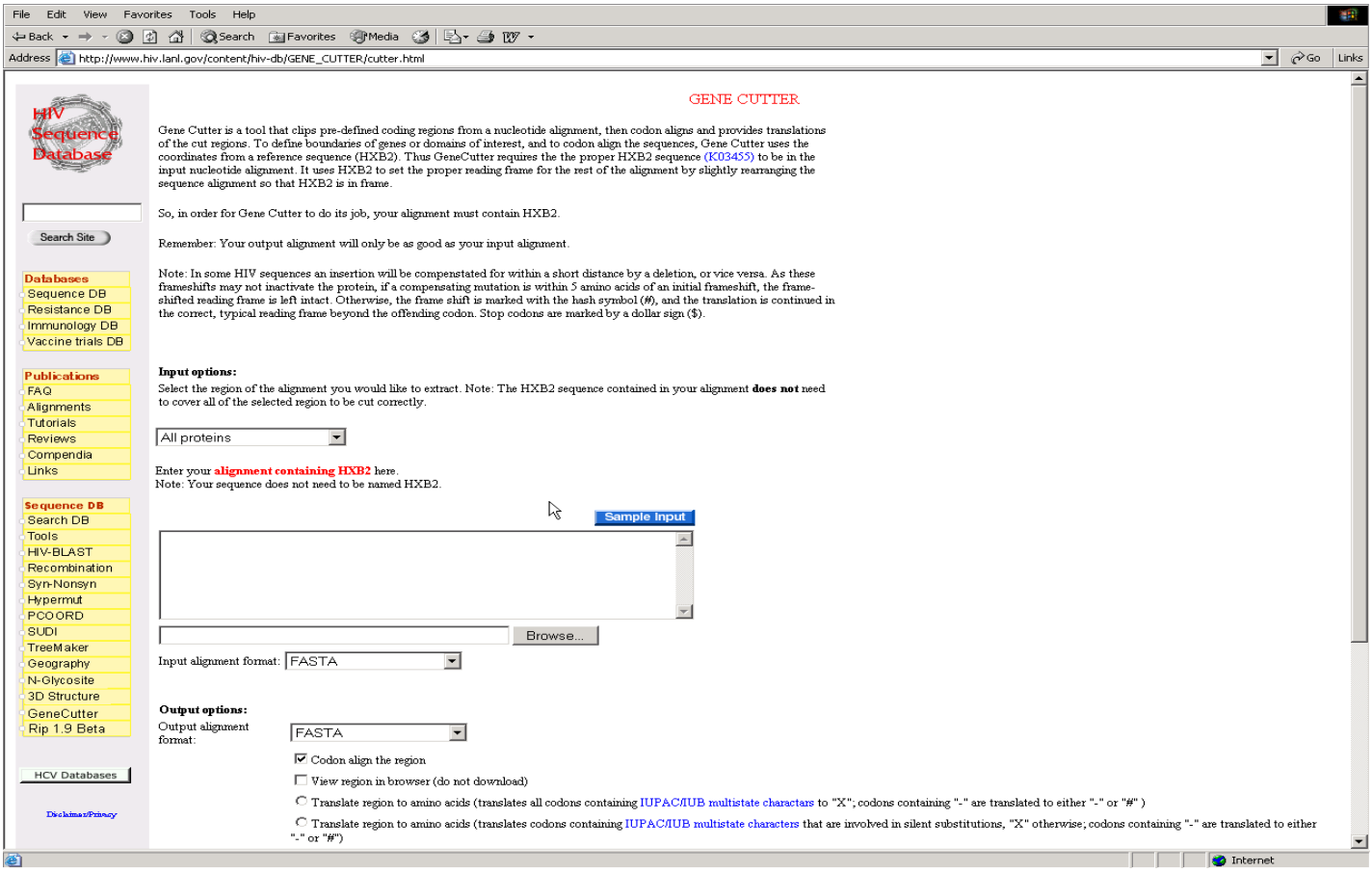
Geography Tool

- Another way to search/download sequences is by geographic region or country
- Results are biased as they show only the sampled individuals, not the true subtype distribution for a region's population
- Results are selectable as in the search interface



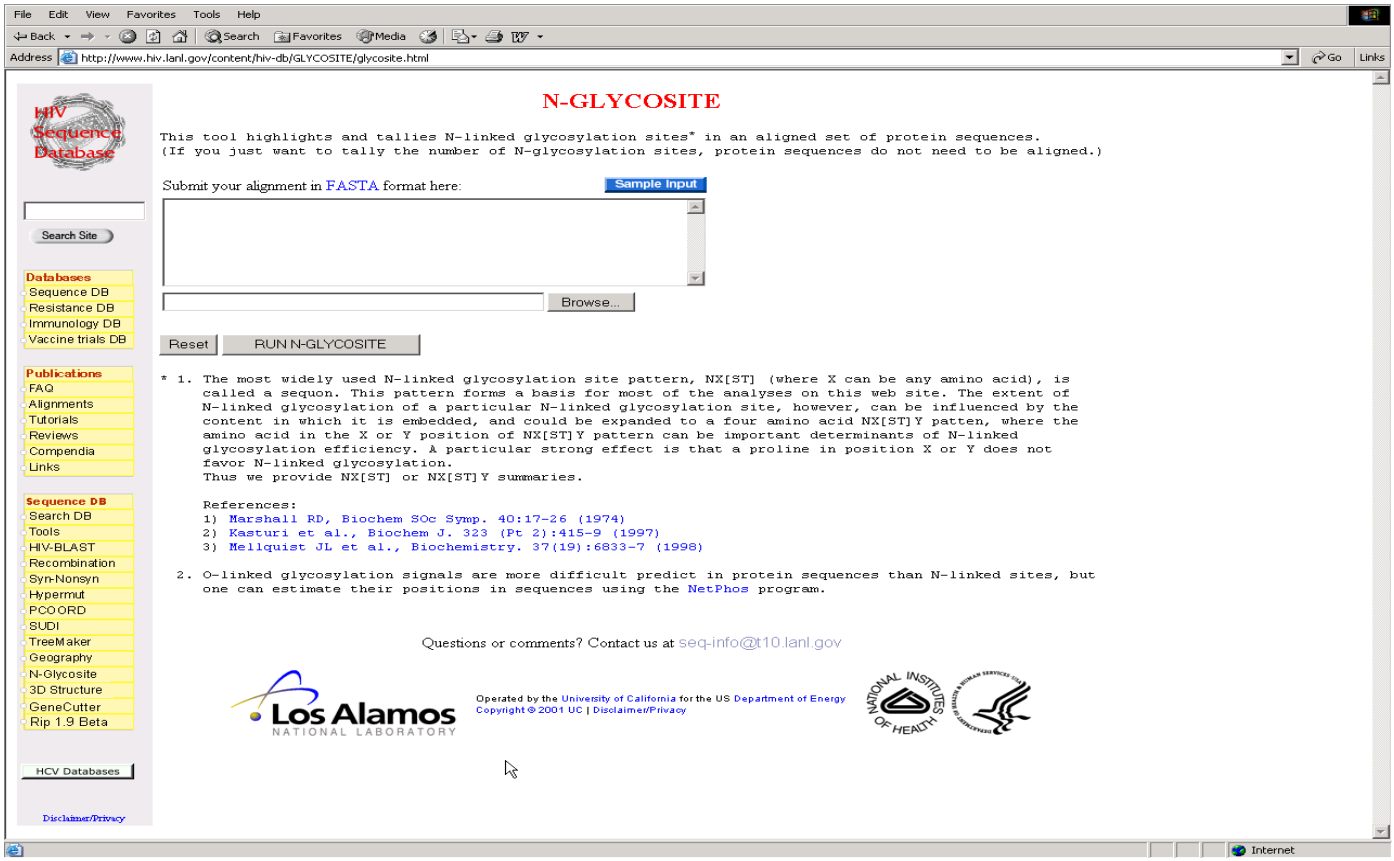
Gene Cutter

- Useful for sequencing labs, particularly for rapid processing of new sets of full length genomes
- Cut out genes and proteins from aligned sets of DNA sequences
- Sequences do not need to be codon aligned – results can be codon-aligned on the fly with generally good results
- Currently, sequence alignments must contain HXB2 as a reference for the program to function



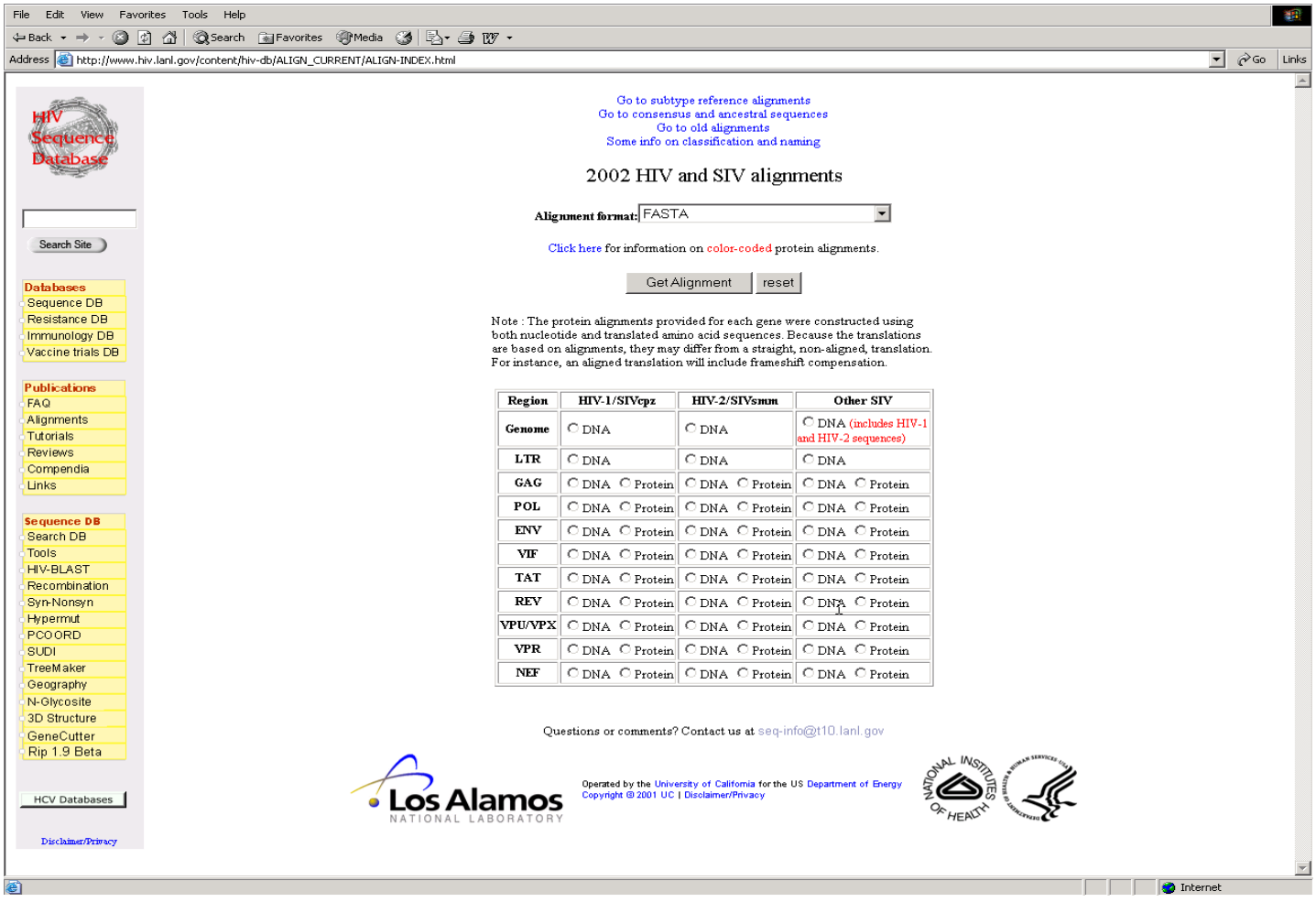
N-GLYCOSITE

- Tracks of patterns of N-linked glycosylation site (N-X-[ST]) change in sequences
- INPUT: A sequence alignment of interest
- OUTPUT:
  - Tallies of numbers of Ngly sites in each sequence
  - Highlighted Ngly sites
  - Graphics illustrating frequency of Ngly patterns in the alignment, and in sub-regions of the alignment
  - Frequencies of different patterns of X and Y in N-X-[ST]-Y



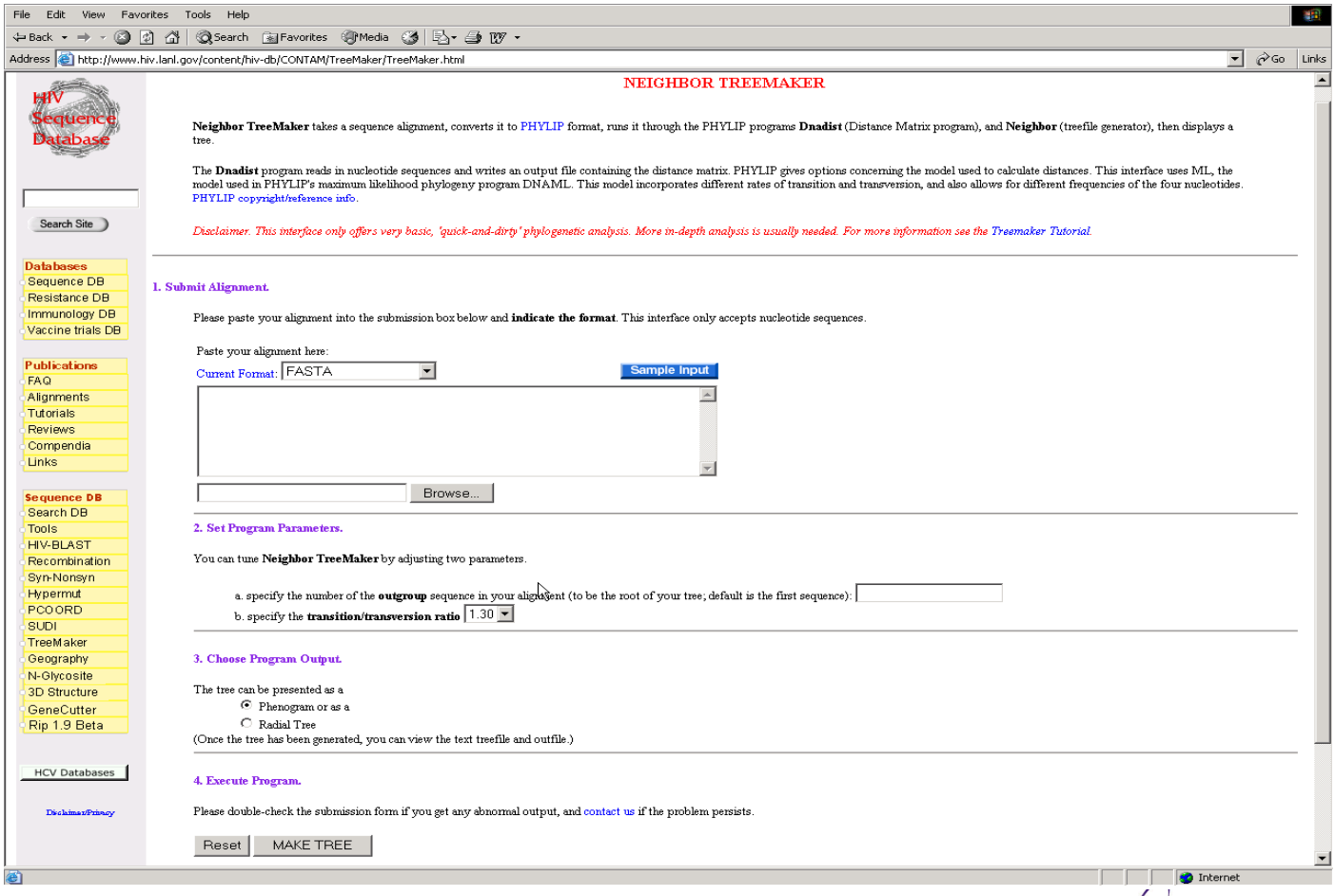
Sequence alignments

- Originally based on iterations of manual and HMM alignments
- Yearly updates using HMM and manual corrections
- Full length genomes updated throughout the year
- Alignments are in reading frame (codon aligned)
- Alignments non-redundant
- Compendia alignments show fewer sequences than web version
- Reference alignments contain up to four representatives
- Protein alignments may contain frameshift compensations
- Subtype consensus with ties resolved, as well as maximum likelihood ancestors, are available for reagent production



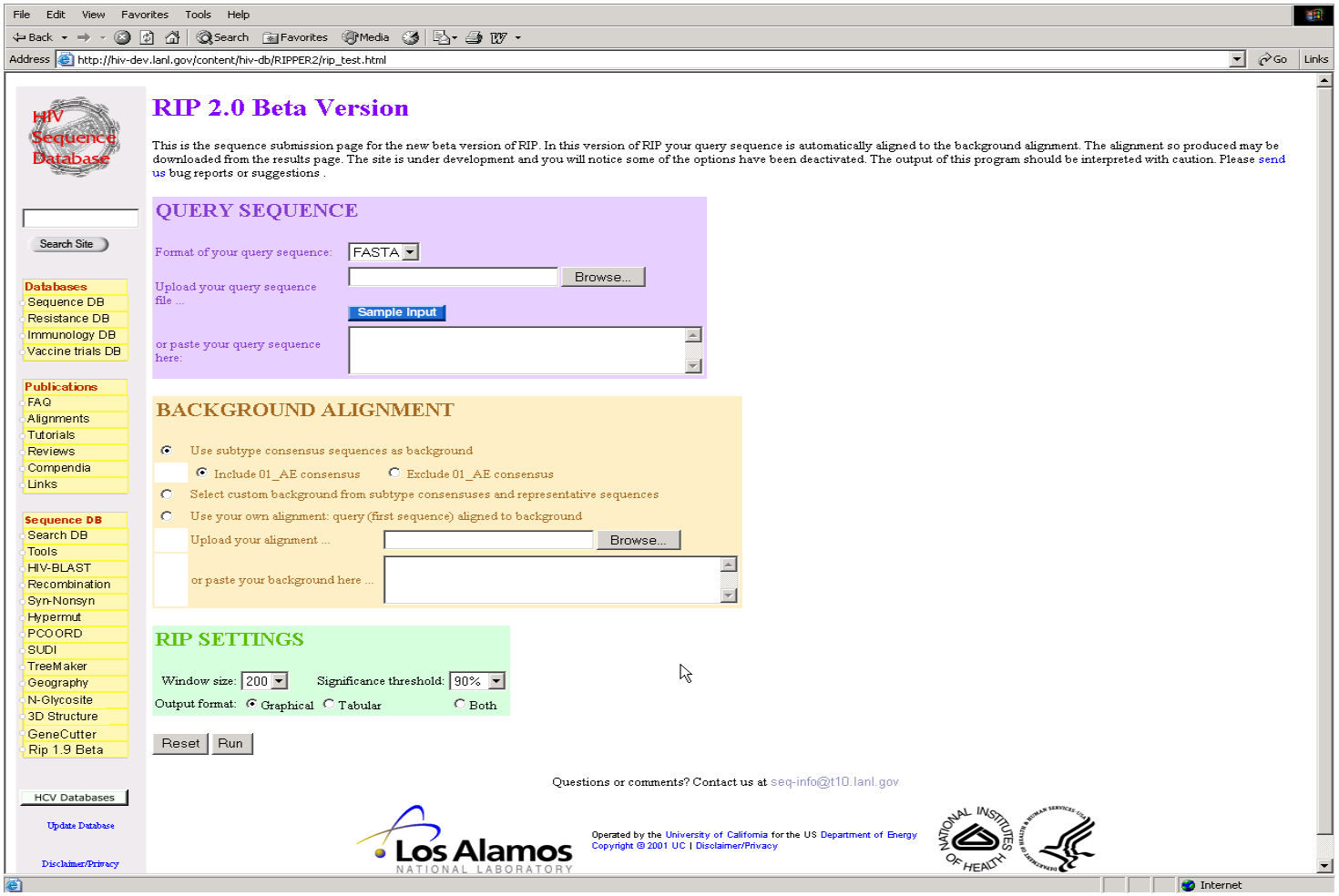
Comparing ”new” and database sequences

- As an example, eight “new” *env* sequences have been aligned to subtype references (A-K) and sequences from previous search (Japanese *env*) using HMM, Genecutter & Se-Al
- TreeMaker produces a Neighbor Joining tree for a “quick-and-dirty” comparison
- TreeMaker is based on DNADIST & NEIGHBOR in the PHYLIP package
- HIV-BLAST is an option for looking for highly similar sequences or possible contamination



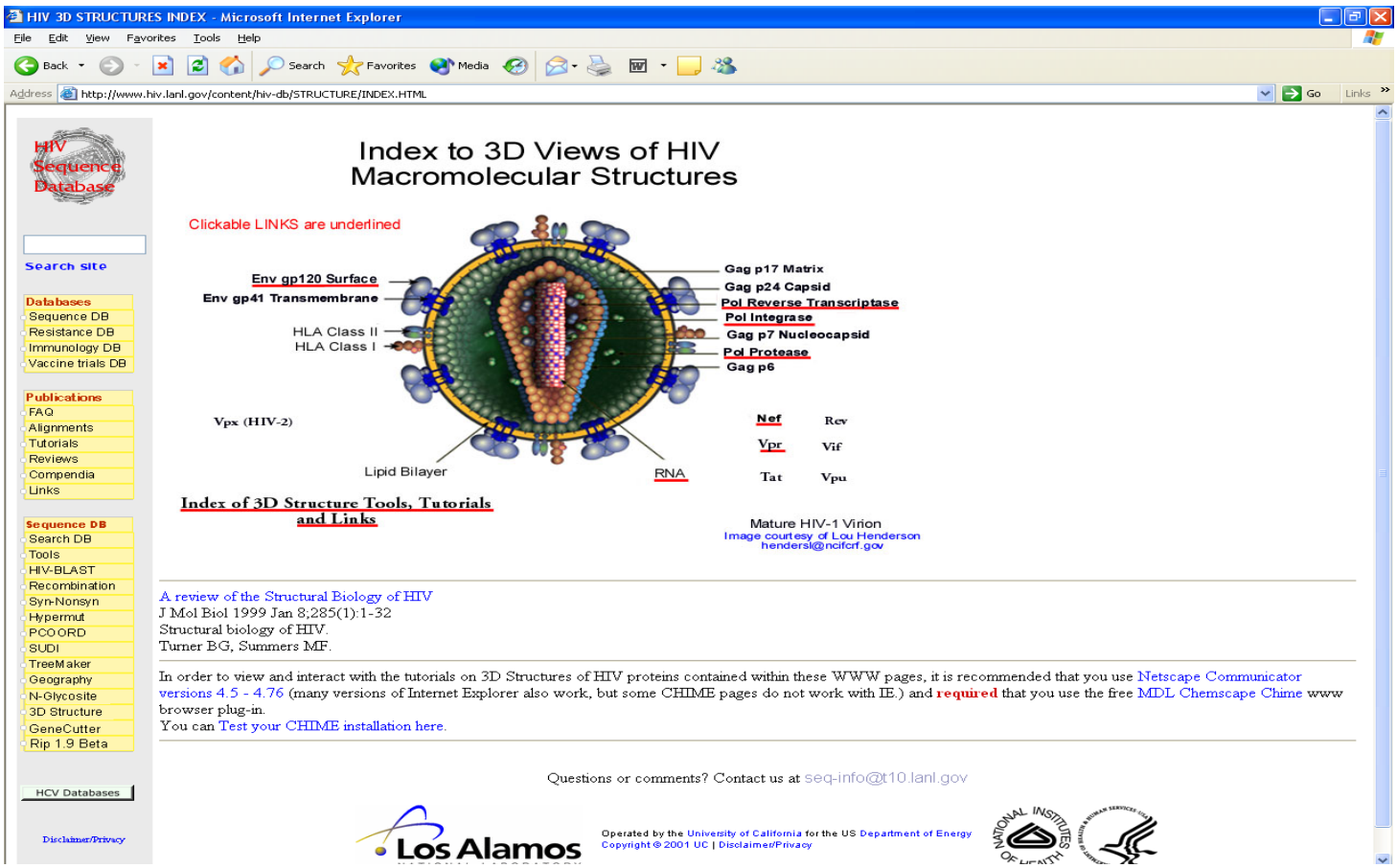
Recombination Analysis

- Many methods and programs exist to investigate potential recombination
  - http://bioinf.man.ac.uk/~robertson/recombination/
- Investigating recombination requires many steps
- A new version of RIP is available at HIV db
  - Automatic alignment
  - Selection of background sequences
  - Different window/gap handling options
  - Graphic & table output



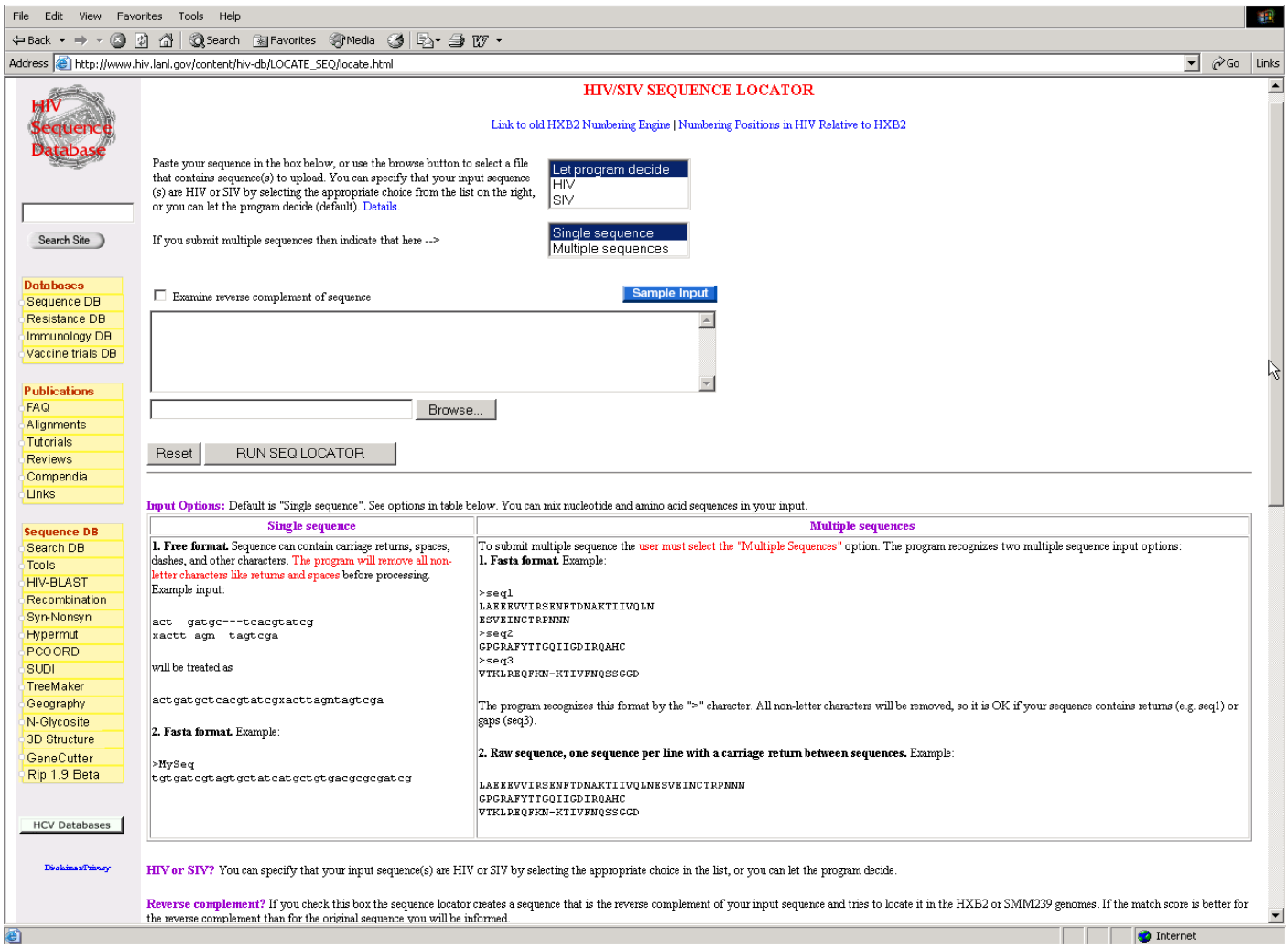
HIV and SIV Protein 3D Structures

- Structures Determined by X-Ray Crystallographic and NMR Methods
  - HIV and SIV Structural proteins (RT, Protease, Env Core)
  - RNA Secondary Structures (TAR, PSI, RRE)
- Models of non-crystalized proteins
  - Env Core with Variable Loops added
- Tutorials and Reviews
  - Env, RT, Protease
- Links to Free 3D Viewer Software
  - CHIME/Protein Explorer
  - RasMol
  - Visual Molecular Dynamics



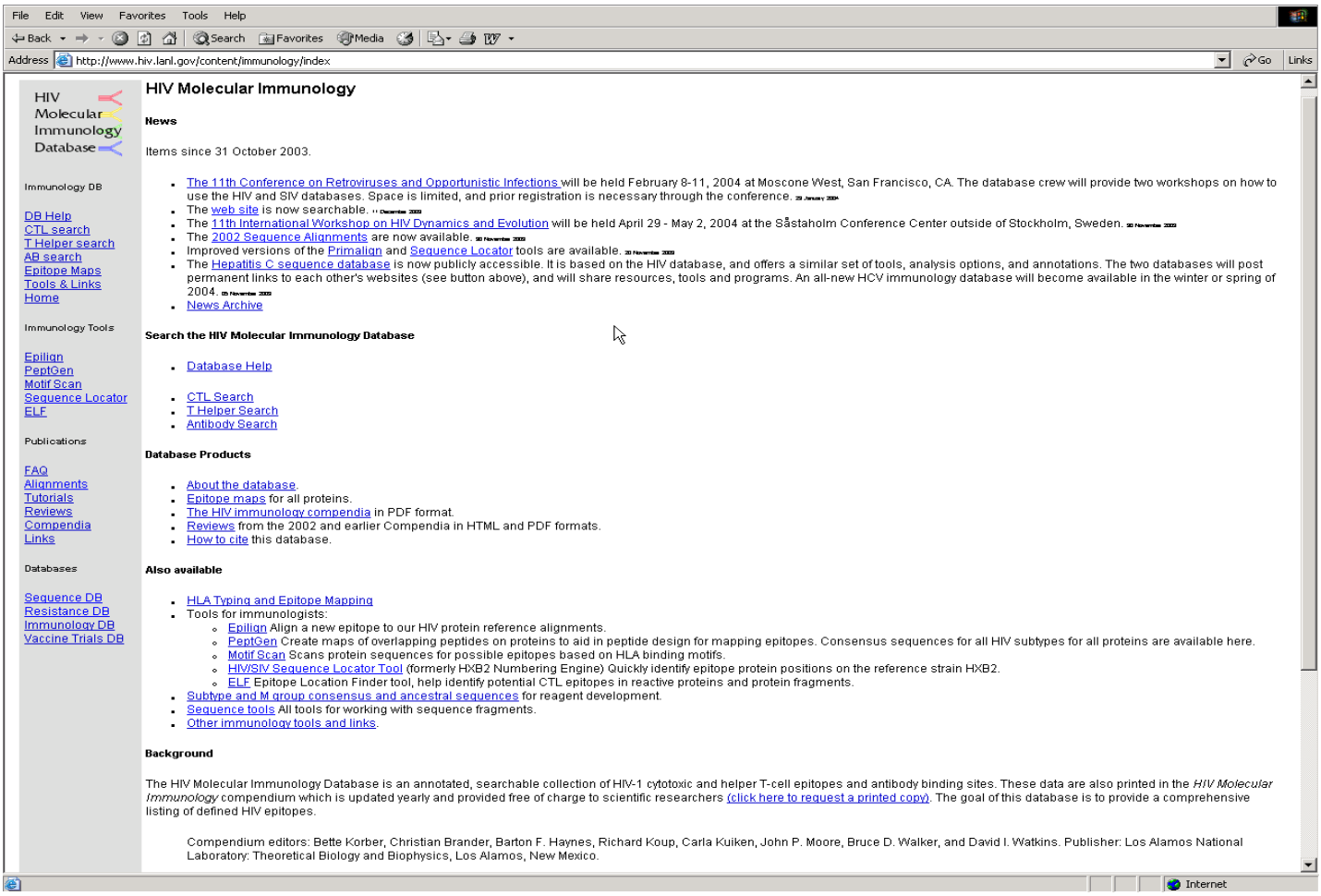
HIV/SIV Sequence Locator Tool

- Rapidly returns position numbers of an HIV or SIV DNA or protein sequence fragment relative to the HXB2r or SMM239 reference strains.
- Such numbers are often included in the literature, and are often incorrect
- Marks the location of the sequence on an HIV map
- For DNA sequences, a translation is provided
- Can be used for input into the search interface, to align a new sequence you have generated with the database set.



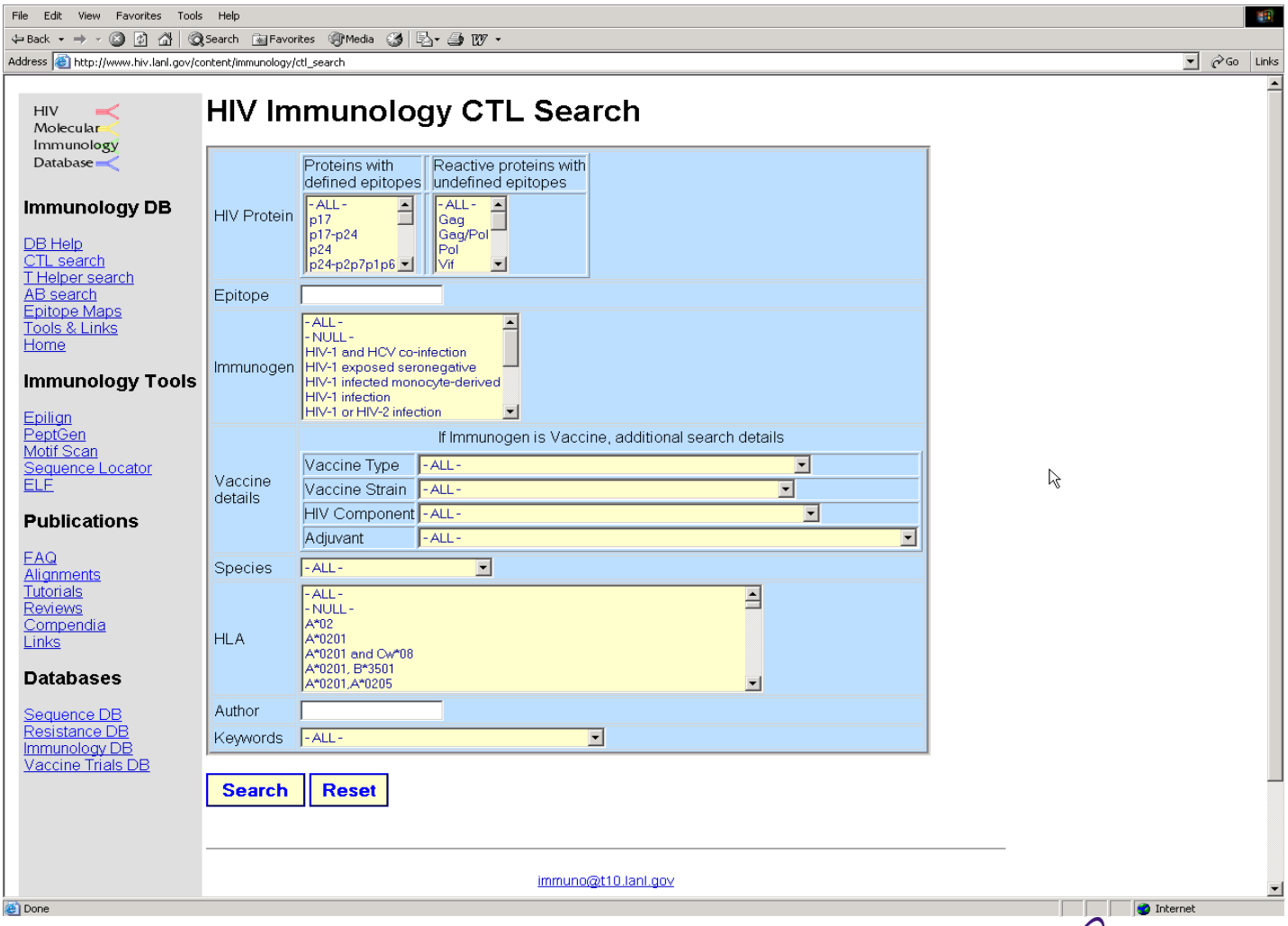
Immunology Database Overview

- HIV T-Cell (CTL, T-helper) and Antibodies (Ab)
- Types of data recorded
  - Epitope sequence and location: HXB2 numbering, subtype
  - Immunogen
  - Host HLA or MHC, and Ab isotype
  - Notes summarize main findings
- Contents: data from 1985 through 2002
  - 2300 CTL entries
  - 600 T-helper entries
  - 1100 Ab entries



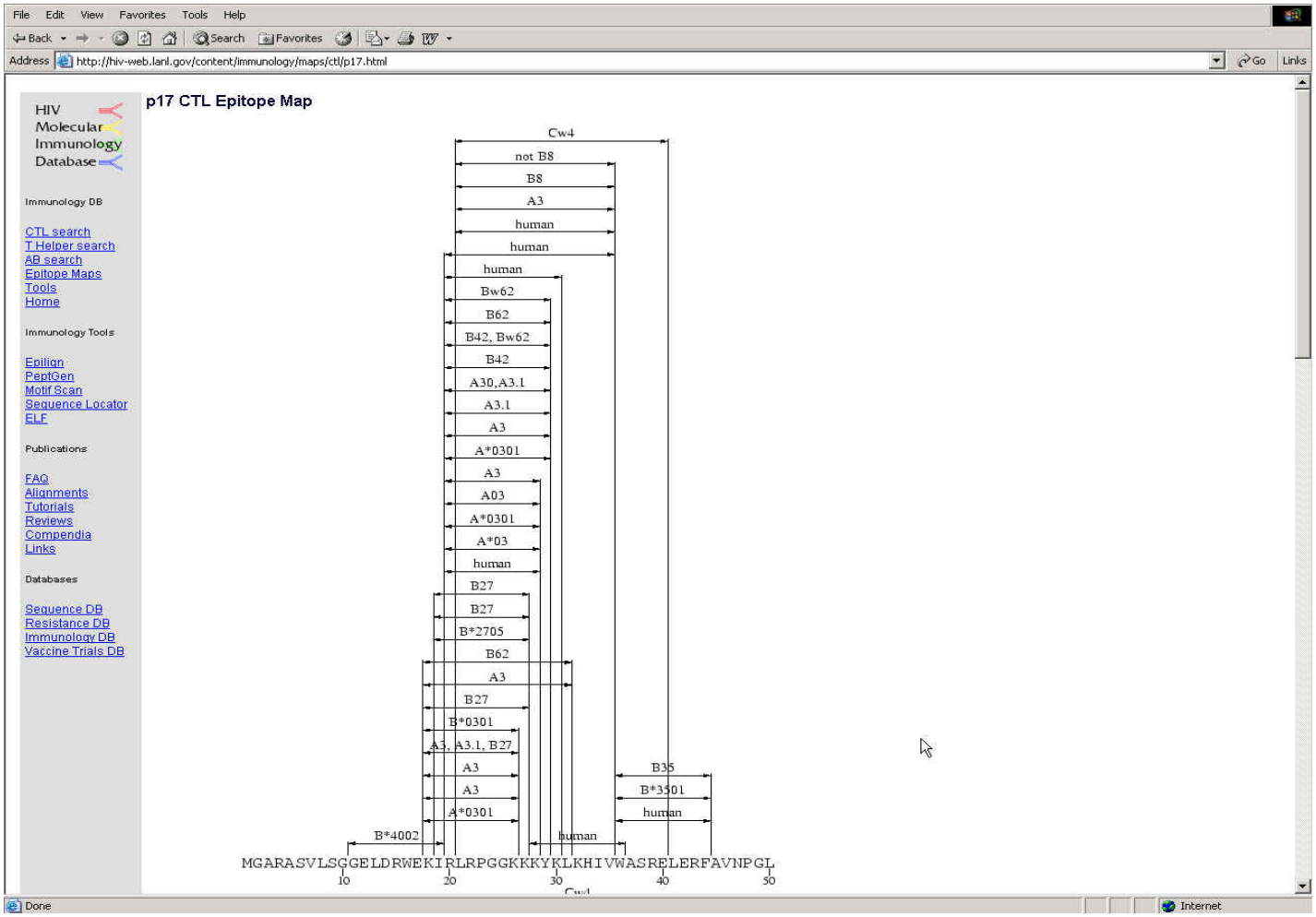
Immunology Database: Search

- T Cells
  - Cytotoxic T Lymphocytes (CTL)
  - Helper T Lymphocytes (T-helper)
  - Biological distinction between CTL and T-helper is not always obvious
  - Organization is identical for CTL and T-helper
  - One reference per entry
- B Cells (Antibodies)
  - One entry for each monoclonal antibodies
  - Many references per entry (up to 100)



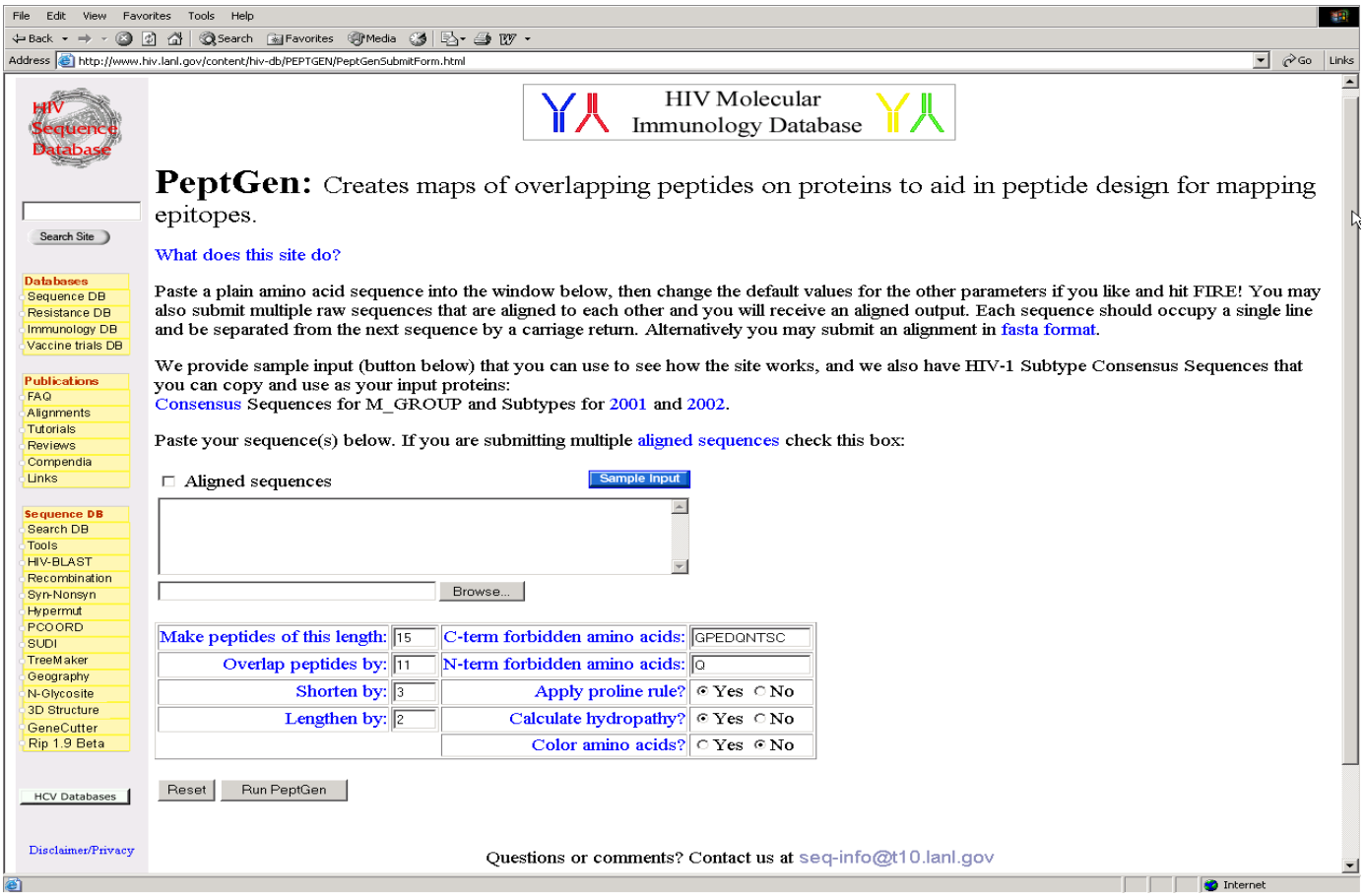
Immunology Database: Additional Information

- All entries for a reference
- Medline links to papers
- Epitope Tables
- Epitope Maps
  - Unique species/HLA for T cell epitopes
  - MAb name, species code for Ab
- Epitope Alignments
  - Extracted from HIV-sequence database, includes subtype, country and year of sampling



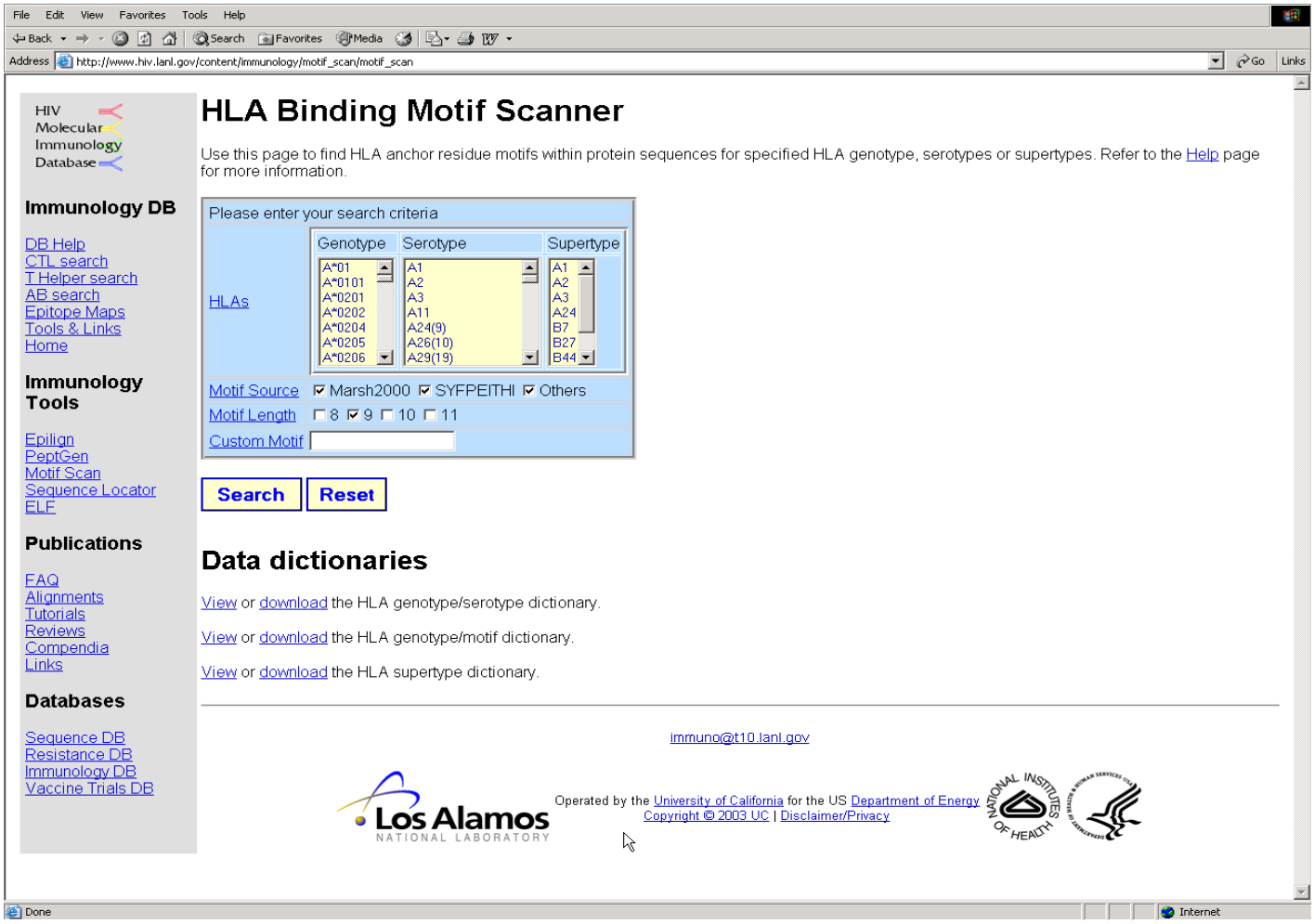
PeptGen

- Creates maps of overlapping peptides on proteins to aid in peptide design for mapping epitopes
- Consensus sequences for all HIV subtypes for all proteins are available
- Use alignments to design comparable sets of peptides (for example, to compare clades)
- INPUT
  - Query sequence or aligned sequences
  - Desired length of peptides, peptide overlap, forbidden C- and N-terminal amino acids
- OUTPUT
  - Maps of overlapping peptides (forbidden amino acids are taken into account)
  - Highlighted forbidden amino acids
  - Hydropathicity scores for the peptides are available



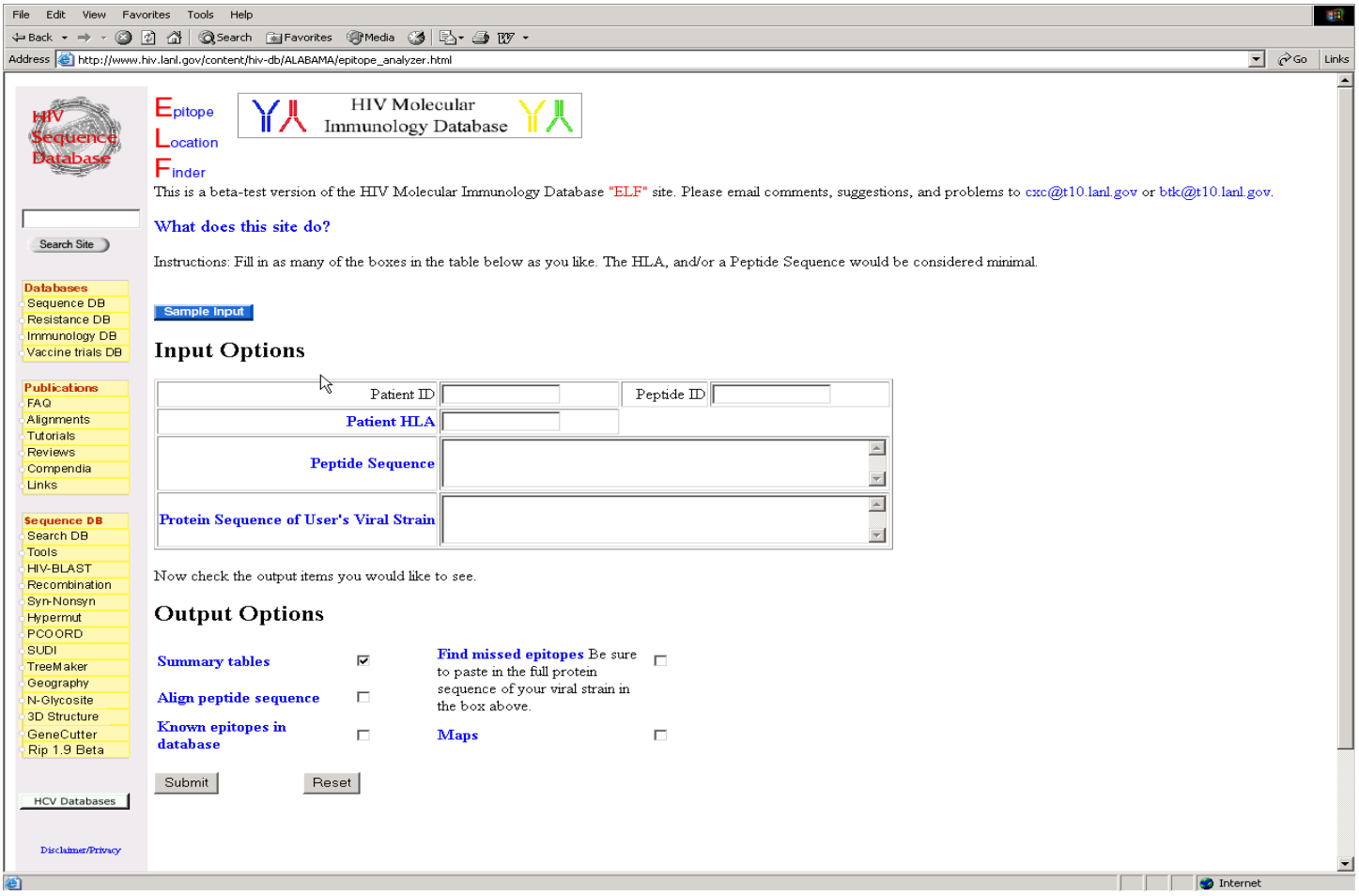
HLA Binding Motif Scanner: MotifScan

- Finds HLA anchor motifs within protein sequences for specified HLA genotypes, serotypes, or supertypes
- HLA anchor motif dictionaries are available on line
- Main motif and supermotif sources:
  - SYPHEITHI Database, Rammensee *et al*. www.syfpeithi.de
  - HLA Facts Book, Marsh *et al*. 2000
  - Sette & Sidney, *Immunogenetics* **50**:201-212, 1999
- INPUT:
  - User defined query sequence or aligned sequences, or reference set
  - Selected HLA anchor motifs, or user defined motif
  - The user defined motif function could be used to search for other patterns of interest in sequences
- OUTPUT:
  - Anchor residue positions are highlighted in the query sequence
  - Potential epitopes and positions are listed
  - Output can be downloaded as text, convenient for further analysis



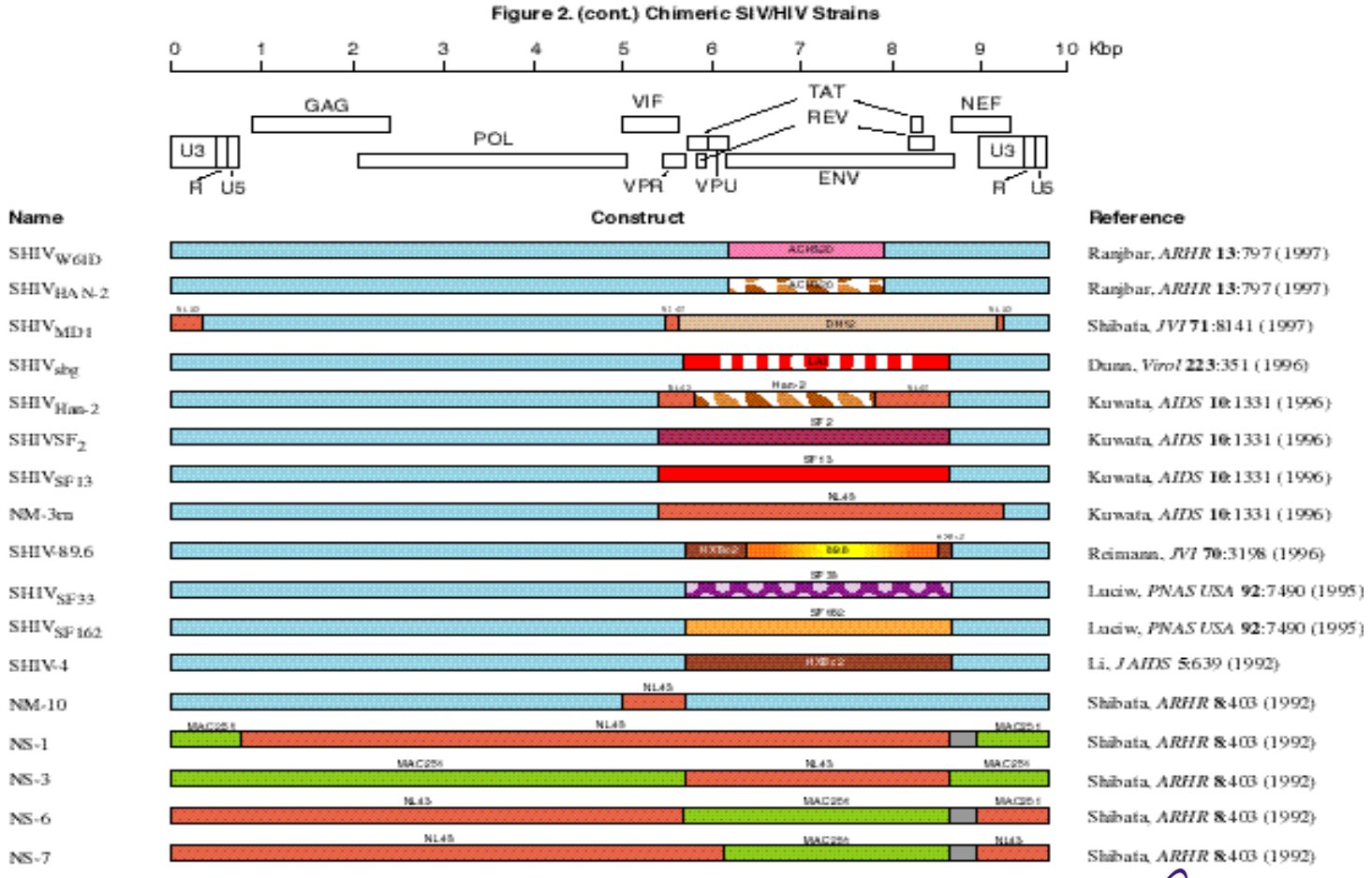
ELF: Epitope Location Finder

- Helps identify potential CTL epitopes in reactive peptides
- INPUT:
  - Reactive peptide sequence
  - Full protein sequence that peptides were based on
  - Patient's HLA information – genotype or serotype
- OUTPUT:
  - If HLA serotype is submitted, associated HLA genotypes are given
  - Potential epitopes in reactive peptides, based on anchor motifs
  - Maps of all HIV epitopes for all HIV proteins, highlighting epitopes that use the patient's HLA presenting molecules
  - Location of the query peptide according to the HXB2 reference strain
  - Alignment of query peptide against reference database alignments
  - All known CTL HIV epitopes contained in the query peptide
  - Epitopes presented by the patient's HLA presenting molecules, that are potentially experimentally missed because of amino acid differences between the previously defined epitopes and the query strain



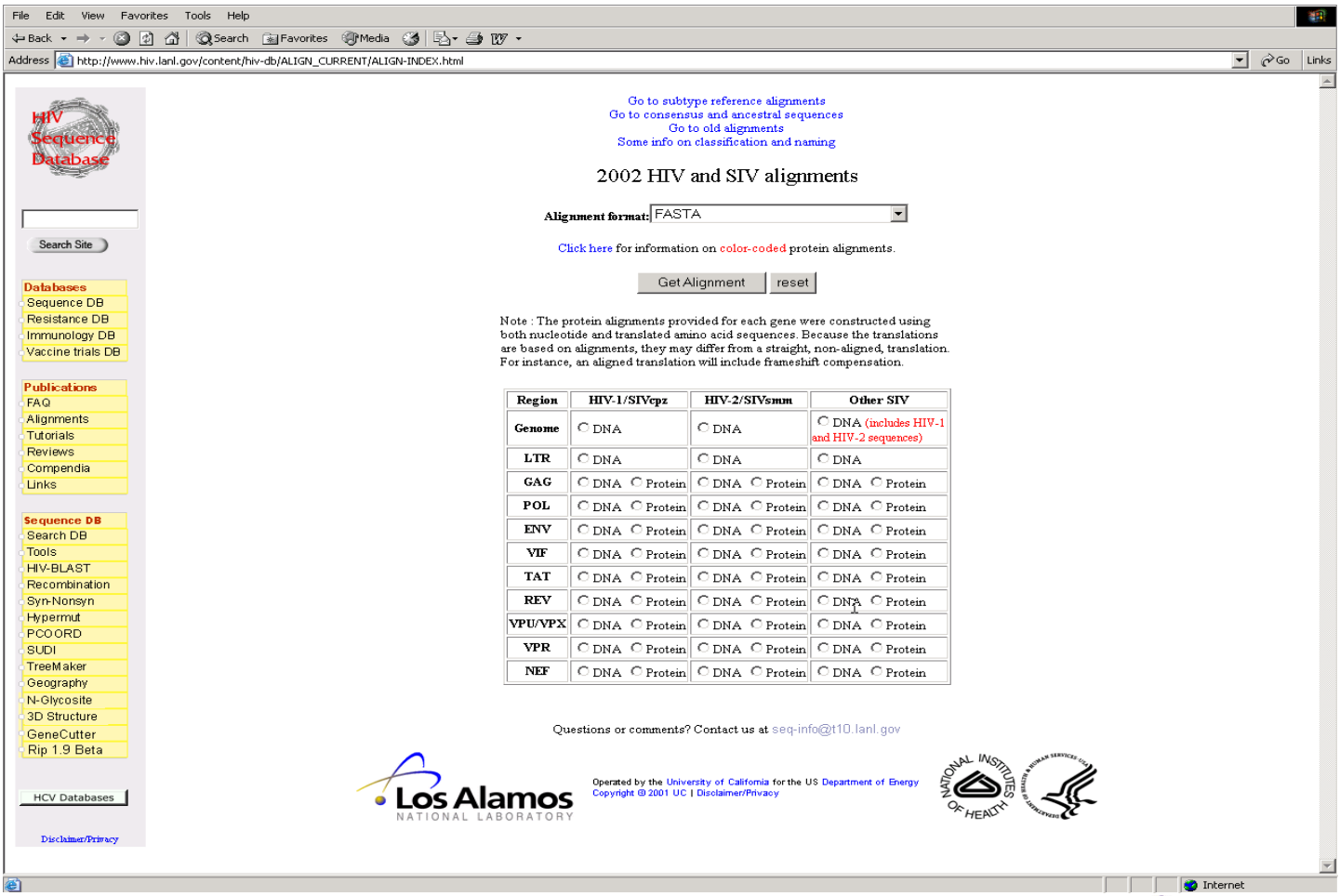
SHIV Maps

- SHIVs are Simian (almost always SIV-SMM or SIV-MAC) / Human (usually HIV-1 subtype B) hybrid infectious clones.
- They are frequently used as vaccine reagents and/or challenge virus.
- Many are derived from clones, but often derived via complex in-vivo passage histories.
- Some have been sequenced, but many are only known via sequence of parental lineages.
- Many authors do not adequately describe the SHIVs they are working with, so we have created a review paper to help describe them.
- http://www.hiv.lanl.gov/content/hiv-db/REVIEWS/VACCINE2001/Vaccine2001.html



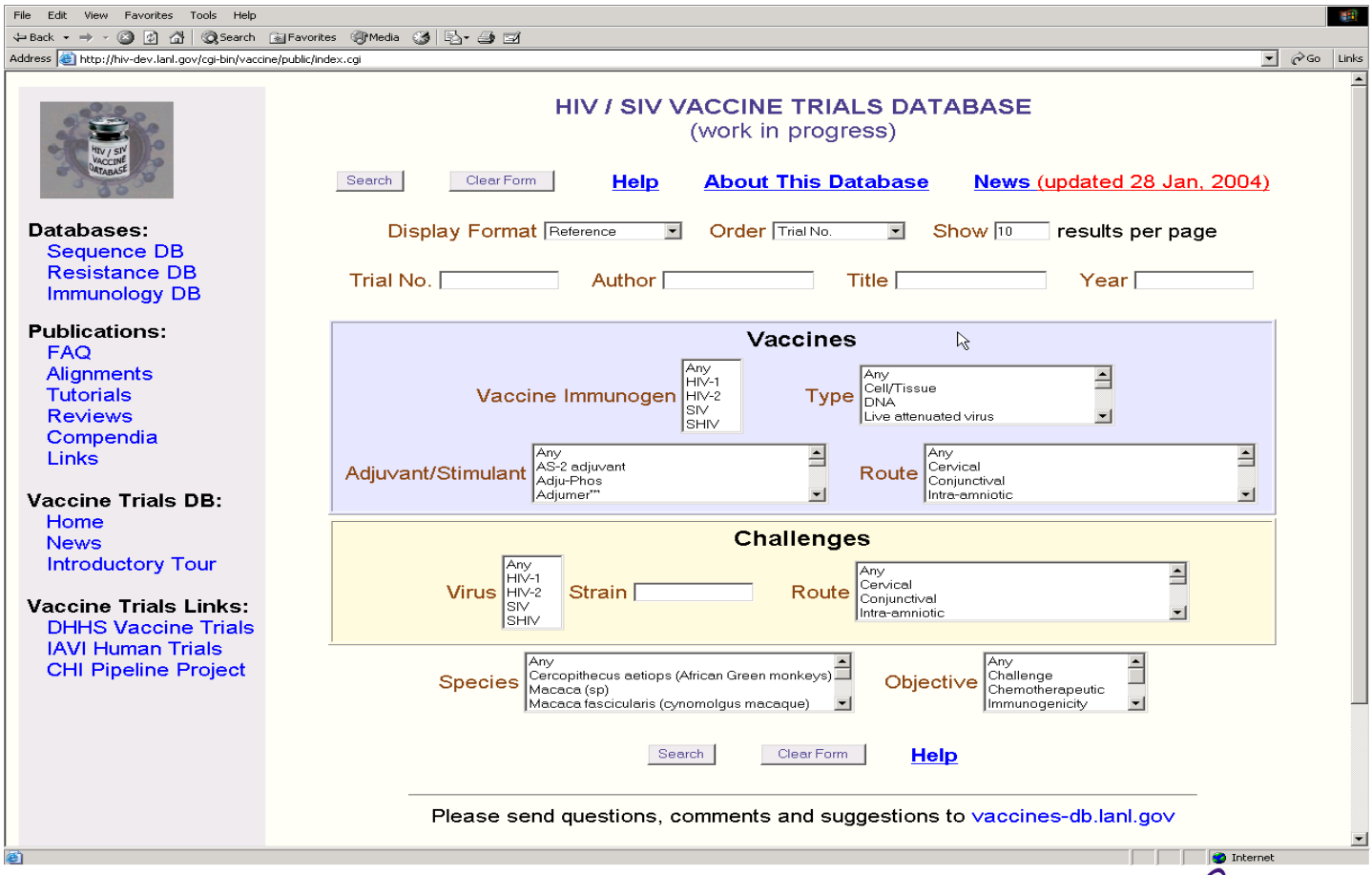
SIV/PIV Alignments

- Any non-human lentivirus is a SIV, not just the SIV-SMM/SIV-MAC group from Sooty mangabeys.
- HIV-1s (M, N and O groups) are related to the SIV-CPZs from the *Pan troglodytes troglodytes* chimps. We describe these alignments as HIV-1/CPZ
- HIV-2s and SIV-MACs are related to SIV-SMMs from Sooty mangabeys. We describe these alignments as HIV-2/SMM
- Dozens of other diverse non-human primates, such as African green monkeys, carry SIVs.
- Alignments of the diverse SIVs, plus HIVs, can help to identify highly conserved codons and other features. We describe these alignments as "other SIV" or HIV-1/HIV-2/SIV.
- http://www.hiv.lanl.gov/content/hiv-db/ALIGN_CURRENT/ALIGN-INDEX.html



Vaccine database

- Under construction
- Non-human primate models
- http://hiv-web.lanl.gov/cgi-bin/vaccine/public/index.cgi
- Aims: To introduce the new database and to interactively retrieve information related HIV/SIV vaccine studies in nonhuman primates
- Search criteria:
  - Vaccine
  - Challenge
  - Adjuvant
  - Objective: Immunogenicity, challenge, etc.
- Sources of Databases
  - LANL
  - EMMS Corporation (Dr Jon Warren)
- Output:
  - Reference
  - Summary
  - Results
  - Example of search
  - Results tabulation



Please leave any comments or suggestions with us: